Large-Scale Orthology Predictions for Inferring Gene Functions Across Multiple Species

Chenggang Yu, Valmik Desai, Nela Zavaljevski, and Jaques Reifman
US Army Medical Research and Materiel Command (MRMC), Biotechnology HPC Software Applications Institute, Telemedicine and Advanced Technology Research Center, Ft. Detrick, MD
{cyu, valmik, nelaz}@bioanalysis.org, jaques.reifman@us.army.mil

Abstract

An effective approach to infer the functions of genes is to use the concept of gene orthology. Because orthologous genes are likely to share similar functions, the functions of genes in an unstudied species can be inferred through the functions of their orthologs in a studied model species. To infer gene functions for a multitude of species, we developed a high-throughput orthology prediction method, termed PhyloTrace. PhyloTrace is both highly accurate and computationally efficient for large-scale applications, having the ability to infer orthologous genes across thousands of species. This is accomplished through three major steps: 1) all-against-all gene comparisons for every pair of genes, 2) pair-wise orthology predictions for every two genomes, and 3) the generation of orthologous clusters that contain orthologous genes across multiple genomes. We employed the previously developed Pipeman parallelization program to break down a set of millions of input sequences into small chunks and then processed them in parallel. We successfully predicted orthologs for over 900 bacterial genomes, achieving a false-positive prediction rate of 2.0%, which was a significant improvement compared with the widely used bidirectional best-hit method, which yielded a false-positive rate of 5.5%.

1. Introduction

Today's high-throughput DNA sequencing technologies are providing complete genome data for a multitude of species at an ever-increasing speed. Genome data for more than one thousand species have been deposited in public databases, such as GenBank, doubling in size approximately every 18 months (Benson, 2008). However, the functions of many genes in these newly sequenced genomes are not known because of the slow and costly experimental studies needed to annotate genomic data, affording the identification of gene function for only a few model organisms. An effective, high-throughput computational approach to infer the functions of genes is to use the concept of orthology, or orthologous genes, which are genes in different species that have evolved from the same ancestral gene in the species' last common ancestor (Koonin, 2005). Because orthologous genes quite likely share similar biochemical functions and biological roles in different species, through orthology, the functions of genes in an unstudied species can be inferred based on their corresponding orthologs in a studied model organism. Using the concept of orthology has helped the understanding of human genes and their mutations associated with cancer by studying their corresponding orthologous genes in mice (Denny, 2000). Moreover, the identification of orthologous genes between emerging and well-studied pathogens is also important for transferring knowledge and gaining insight into the pathogenicity of emerging pathogens. In addition, as part of the US Department of Defense (DoD) biological defense programs, the identification of orthologous genes conserved across related pathogenic organisms is currently being pursued as a strategy for identifying broad-spectrum drug targets.

The growing needs for orthology information and the rapid accumulation of genomic data call for orthology prediction methods that are both highly accurate, which is critical for function inference, and computationally efficient, which is required for large-scale, high-throughput applications, e.g., ortholog predictions for thousands of species, each consisting of thousands of genes (Gabaldon, 2009). Some of the existing methods are computationally efficient but not very accurate, whereas other methods are accurate but not efficient, even those making use of high performance computing (HPC) resources (Gabaldon, 2009). We developed a high-throughput, HPC-based

978-0-7695-4392-5 2010

DOI 10.1109/HPCMP-UGC.2010.40

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
JUN 2010		**00-00-2010 to 00-00-2010**

4. TITLE AND SUBTITLE
Large-Scale Orthology Predictions for Inferring Gene Functions Across Multiple Species

5a. CONTRACT NUMBER
5b. GRANT NUMBER
5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

5d. PROJECT NUMBER
5e. TASK NUMBER
5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command,Biotechnology High Performance Computing Software Applications Institute,Telemedicine and Advanced Technology Research Center,Fort Detrick,MD,21702

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)
11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
2010 DoD High Performance Computing Modernization Program Users Group Conference, 14-17 Jun, Chicago, IL.

14. ABSTRACT
An effective approach to infer the functions of genes is to use the concept of gene orthology. Because orthologous genes are likely to share similar functions the functions of genes in an unstudied species can be inferred through the functions of their orthologs in a studied model species. To infer gene functions for a multitude of species, we developed a high-throughput orthology prediction method, termed PhyloTrace. PhyloTrace is both highly accurate and computationally efficient for large-scale applications, having the ability to infer orthologous genes across thousands of species. This is accomplished through three major steps: 1) allagainst- all gene comparisons for every pair of genes, 2) pair-wise orthology predictions for every two genomes and 3) the generation of orthologous clusters that contain orthologous genes across multiple genomes. We employed the previously developed Pipeman parallelization program to break down a set of millions of input sequences into small chunks and then processed them in parallel. We successfully predicted orthologs for over 900 bacterial genomes, achieving a falsepositive prediction rate of 2.0%, which was a significant improvement compared with the widely used bidirectional best-hit method, which yielded a falsepositive rate of 5.5%.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**4**	

orthology prediction program, termed PhyloTrace, which addresses these two limitations. PhyloTrace is more accurate than the widely used existing methods for large-scale applications, whereas its computation time is equivalent to other computationally efficient methods. In this paper, we describe the PhyloTrace program and its parallelization on a Linux computer cluster. We show that PhyloTrace provides accurate gene orthology prediction for a large number of species in a computationally efficient manner.

2. Methods

PhyloTrace predicts orthologs for a group of genomes through three major steps (Figure1): 1) all-against-all gene sequence comparisons for every pair of genes in the input genomes, 2) pairwise orthology predictions for every two genomes, and 3) the generation of orthologous clusters that contain orthologous genes across multiple genomes.

In the first step, we used the BLAST program (Altschul, 1990) to calculate the sequence similarity for every two genes in the input genomes by comparing each gene with all other genes in an indexed database of the input genomes. Next, a post-processing program (Postproc) searched the BLAST outputs for all pairs of genes that are bidirectional best-hits (BBH). Two genes, ***a*** and ***b***, from two genomes, ***A*** and ***B***, respectively, form a BBH pair if gene ***a*** has the highest sequence similarity with gene ***b*** compared with all genes in ***A,*** and gene ***b*** also has the highest sequence similarity with gene ***a*** compared with all genes in ***B***. Genes that form BBH pairs are considered to be orthologs in many applications (Mushegian, 1998), although this method yields a significant amount of false predictions (Koonin, 2005).

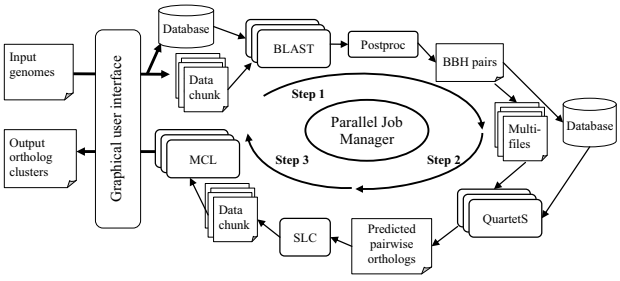

Figure 1. Overall architecture of the high-throughput PhyloTrace orthology prediction program. BBH, bidirectional best hits; SLC, Single Linkage Cluster; MCL, Markov Cluster Algorithm.

In the second step, we initially inferred the BBH pairs identified in the first step to form orthologs. Then, we used the in-house-developed QuartetS algorithm to identify genes that form BBH pairs, but are not orthologs. Such putative non-orthologs include those genes that form a BBH pair where two genes in the pair are found to form BBH pairs with different genes in other genomes. In such cases, QuartetS analyzes the evolutionary relationship of the two genes and infers possible evolutionary events that can determine if they should be considered to be orthologs. Through the elimination of possible non-orthologous BBH pairs, fewer, but more accurate, orthology predictions are obtained in the second step. The accuracy of QuartetS can be adjusted by changing a cut-off value that determines to what extent two genes that form BBH pairs should be rejected as orthologs.

In the third step, we predicted orthologs among the multiple input genomes by clustering the pairwise orthologs predicted in the second step. This is achieved by forming orthologous clusters, where, in theory, every two genes in a cluster are orthologs, and all of them share similar functions. Although an orthologous cluster can be created if every two genes in a cluster have been predicted to be orthologs by pairwise orthology prediction, in practice, some genes may not be orthologous to every gene in the cluster. In such case, additional analysis is required to determine whether to exclude a gene that is orthologous to some but not all of the other genes in the cluster. This problem can be transformed into a standard network clustering problem, where genes are represented as vertices of a graph, and their pairwise orthology relationships as edges. Two programs, Single Linkage Cluster (SLC) and Markov Cluster Algorithm (MCL) (Dongen, 2008), were implemented in PhyloTrace to cluster genes into orthologous clusters. The SLC program searches isolated sub-networks that are not connected to other parts of the network, whereas the MCL program searches clusters in each sub-network. Through this two-step clustering, orthologous clusters are created as the final output of PhyloTrace.

To effectively predict orthologs for a large number of genomes, we parallelized each of the three steps of PhyloTrace using a data-driven strategy based on a previously developed parallel job management program, Pipeman (Yu, 2007). In the first step, we segmented the input genomes consisting of millions of genes into small non-overlapping chunks of hundreds of genes, where the BLAST program runs in parallel on multiple nodes simultaneously to process each chunk of the data. The Pipeman program controls the parallel execution of the BLAST program using the Message-Passing Interface (MPI) to monitor and manage the execution of programs on multiple nodes. When one node running BLAST finishes one chunk of the data, Pipemen sends it another chunk. When all chunks are completed, Pipeman starts the Postproc program, which collects all results and searches for BBH pairs, saving the results in multiple files and in a database as well. In the second step, Pipeman

starts QuartetS on multiple nodes to compute orthologs in parallel for each pair of genomes. Then, in the third step, the first clustering program, SLC, creates multiple sub-networks. Finally, Pipeman starts the MCL program on multiple nodes to cluster the sub-networks in parallel and create orthologous clusters.

PhyloTrace is deployed at the MANA Linux computer cluster at the Maui HPC Center (MHPCC). It can be accessed through a Web-based graphical user interface (GUI) developed using the User Interface Toolkit. Through the GUI, users can upload their genomes onto the computer cluster and run PhyloTrace to predict orthologs. When the prediction is finished, they can view the results through the GUI and download them onto local machines.

3. Results

We used PhyloTrace to predict orthologs for 949 prokaryotes, which include all prokaryotes whose genomes had been deposited in GenBank as of September 2009 (Benson, 2008). These predictions, which took about 100 hours using 300 cores, would have taken >2 years if run on a single-core desktop computer. Figure 2 shows the speed-up curve for PhyloTrace obtained on MANA. The parallelization yields a nearly ideal speed-up for up to 60 cores, after which the computational efficiency tails off primarily because of the BLAST program, which causes an input/output (I/O) bottleneck when saving a large amount of output data files on shared hard disks. Nevertheless, PhyloTrace still achieves a 200-fold speedup with 300 cores and yields a 75% efficiency, or better, for up to 250 cores, which is more than adequate for making large-scale prediction of orthologs in a reasonable time.

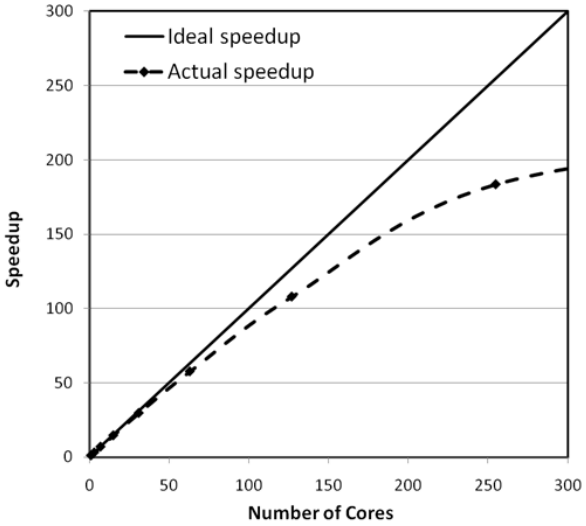

Figure 2. The speedup curve for PhyloTrace, showing that it can achieve 75% efficiency, or better, for up to 250 cores

We validated PhyloTrace's orthology predictions based on a large-scale study using protein functions annotated in the KEGG database (Kanehisa, 2010) as a gold standard. For two genes predicted as orthologs by PhyloTrace, we extracted their functions as annotated in the KEGG database. If the two genes had the same annotated function, the orthology prediction was counted as a true-positive; if they had different annotations, the prediction was counted as a false-positive; and if either one of the two genes had no annotation, the orthology prediction was counted as unevaluated. The performance of PhyloTrace was measured in terms of the prediction rate, which is defined as the fraction of all predictions over the evaluated BBH pairs, and the false-positive rate, defined as the fraction of false-positive predictions over all evaluated PhyloTrace predictions.

Figure 3 shows such validation results for 949 prokaryotes from GenBank. PhyloTrace can yield results with different accuracy, as shown by the solid squares in Figure3, which reflect different cut-off values in the QuartetS definition of orthologs. The figure indicates that the false-positive rate of PhyloTrace can be considerably reduced from ~5.5% (for a cut-off that is equivalent to the BBH method) to <2%. As expected, there is a trade-off between the prediction rate and the false-positive rate, with the prediction rate decreasing when the accuracy increases. Although PhyloTrace predicted fewer orthologs compared with the BBH method, it achieved a >50% reduction in the false-positive rate by reducing the prediction rate by <20%, a significant improvement for applications that require high-quality orthology predictions.

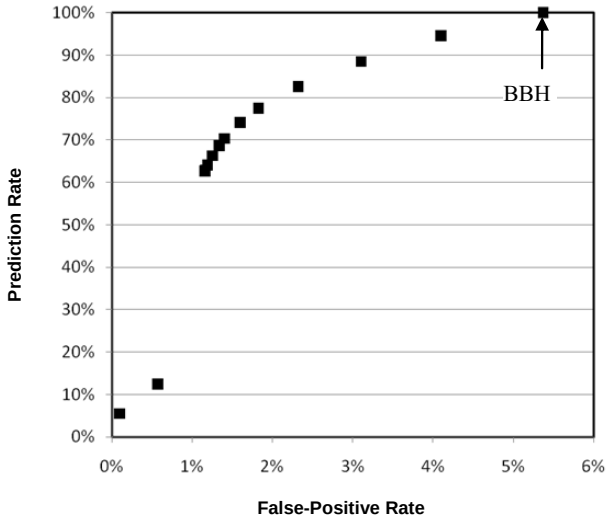

Figure 3. The validation of PhyloTrace for the predictions of ~900 prokaryotes, showing a considerable improvement over the BBH method

4. Conclusions

The developed PhyloTrace method for gene orthology prediction is both accurate and computationally efficient, providing the means for high-throughput genome annotation of newly sequenced pathogens of military relevance and their near-neighbors. PhyloTrace is fully parallelized and operational at the MANA Linux computer cluster at the MHPCC, where we performed orthology prediction for >900 prokaryotic genomes (consisting of ~4 million genes) in the GenBank database in ~100 hours using 300 computing cores. The validation results indicated that, compared with the widely used BBH method, PhyloTrace reduced the false-positive rate of the predictions from 5.5% to 2.0%, with a ~20% reduction in the prediction rate. The user, however, has the ability to trade-off accuracy versus prediction rate for specific applications by setting different cut-off values in the algorithm. Given the unprecedented throughput and cost reduction afforded by the next-generation genome sequencing technologies currently available, we expect gene orthology prediction methods, such as PhyloTrace, to become a vital component of future comparative genomics studies.

5. Significance to DoD

Scientists at the US Army Medical Research Institute of Infectious Diseases are using PhyloTrace in multiple projects sponsored by the Defense Threat Reduction Agency (DTRA) to predict orthologous genes that are conserved across a multitude of bacterial pathogens in search of broad-spectrum drug targets. In addition, PhyloTrace is being used by our laboratory, the DoD Biotechnology HPC Software Applications Institute, in another DTRA-sponsored project to identify effector proteins in *Burkholderia mallei*, the causative agent of glanders.

Disclaimer

The opinions and assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the views of the US Army or of the US Department of Defense.

Acknowledgments

This work was sponsored by the US DoD High Performance Computing Modernization Program (HPCMP), under the High Performance Computing Software Applications Institutes (HSAI) Initiative.

References

Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman, "Basic local alignment search tool." *J. Mol. Biol.*, 215, pp. 403–410, 1990.

Benson, D.A., I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, and D.L. Wheeler, "GenBank." *Nucleic Acids Res*, 36, D25–D30, 2008.

Denny, P. and M.J. Justice, "Mouse as the measure of man?" *Trends in Genetics*, 16:7, pp. 283–287, 2000.

Dongen, S., "Graph clustering via a discrete uncoupling process." *Siam Journal on Matrix Analysis and Applications*, 30-1, pp. 121–141, 2008.

Gabaldon, T., C. Dessimoz, J. Huxley-Jones, A.J. Vilella, E.L. Sonnhammer, and S. Lewis, "Joining forces in the quest for orthologs." *Genome Biol*, 10:403, 2009.

Kanehisa, M., S. Goto, M. Furumichi, M. Tanabe, and M. Hirakawa, "KEGG for representation and analysis of molecular networks involving diseases and drugs." *Nucleic Acids Res*, 38, D355–D360, 2010.

Koonin, E.V., "Orthologs, paralogs, and evolutionary genomics." *Annu Rev Genet*, 39, pp. 309–338, 2005.

Mushegian, A.R., J.R. Garey, J. Martin, and L.X. Liu, "Large-scale taxonomic profiling of eukaryotic model organisms: a comparison of orthologous proteins encoded by the human, fly, nematode, and yeast genomes." *Genome Res*, 8, pp. 590–598, 1998.

Yu, C. and P. Wilson, "A tool for creating and parallelizing bioinformatics pipelines." *Proceedings of the 2007 DoD High Performance Computing Modernization Program Users Group Conference*, pp. 417–420, 2007.